Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 14 (2016) 4122 – 4129

Transportation Research Procedia

www.elsevier.com/locate/procedia

6th Transport Research Arena April 18-21, 2016

The identification of patterns of interurban road accident frequency and severity using road geometry and traffic indicators

Bahar Dadashova [a,]*, Blanca Arenas Ramírez [b], José Mira McWilliams [b], Francisco Aparicio Izquierdo [b]

[a]*Transportation Institute, Texas A&M University, 2935 Research Parkway, College Station, TX77843-3135, USA*
[b]*University Institute of Automobile Research (INSIA), Technical University of Madrid, José Gutiérrez Abascal 2, 28006 Madrid, Spain*

Abstract

This paper is focused on the effect of road geometry, and other accident causing conditions, on the binary response variable road accident severity. The data is collected from two interurban routes in Spain (Madrid-Irún and Barcelona-Almeria) and covers a 3 year period (2010-2012). Data mining techniques were applied for the treatment and combination of two databases for road accident associated factors and geometry standards respectively. The effect of the influential factors on road accident severity was estimated through a non-parametric statistical methodology, random forests. Several standards of the road geometry design were found to have a significant effect on the road accident severity.

Peer-review under responsibility of Road and Bridge Research Institute (IBDiM)

Keywords: Road accidents; victim severity; classification trees; random forest; data mining; freight transport

Corresponding author. Tel.: +1-(979)-845-7415; fax: +1-(979)-845-6006.
E-mail address: bahar.d@tamu.edu

Peer-review under responsibility of Road and Bridge Research Institute (IBDiM)
doi:10.1016/j.trpro.2016.05.383

1. Introduction

Vehicle accidents are complex processes which can be explained as a consequence of different influential factors. Crash data have been mainly analyzed through parametrical statistical tools since the accident occurrence is assumed to follow a given distribution (Poisson, negative-binomial, state-space models, etc.). Application of non-parametric techniques such as classification and regression trees (CART) is a relatively new field in vehicle accident analysis. The list of studies includes Kuhnert et al. (2000), Karlaftis and Golias (2002), Chang and Chen (2005), Abdel-Aty et al. (2008), Harb et al. (2009), Das et al. (2009, 2011), etc.

In this article the road accident severity in Spain's two most frequently used and crowded routes (Madrid-Irún and Barcelona-Almeria) is being analyzed. The accidents that have resulted in at least one injury or one fatality have been recorded during the course of three years (2010-2012). In order to explain the severity of these accidents the road geometry design standards were used. Other significant factors for explaining the accident severity are road type, traffic density at the time of the crash, surface conditions, alignment and visibility obstruction. Thus, given the nature of the variables the data obtained from two different sources: database containing the road geometric design of the routes and database containing the information on accidents and conditions under which the accidents took place. These two databases were merged using data mining tools. In this work 17 variables form the unified database have been selected for random forest analysis.

The rest of this article is structured as follows: in section 1 the data is introduced. In section 2 a brief introduction to random forests is presented. Section 3 presents the results and the discussion. The paper ends with conclusions and references.

2. Data

Two main routes are the focus of the study, those connecting Madrid (central region) with Irún (north) and Almería (south-east) with Barcelona (north-east) as described in Figure 1. This selection coincides with a section of two International Rail Freight Corridors across Spain, as part of TEN-T (Trans-European Transport Network) Program. The former route coincides with part of the International Rail Freight Corridor 4, which runs from Lisbon, Sines and Leixões (Portugal) to Algeciras, Madrid, Bilbao, San Sebastían and Irún (Spain) and all the way up through Paris and into north-eastern France, while the latter follows the Spanish share of the Rail Freight Corridor 6, which runs along the south of Europe from Almeria and Madrid in Spain to Záhony in Hungary, crossing France, Italy and Slovenia.

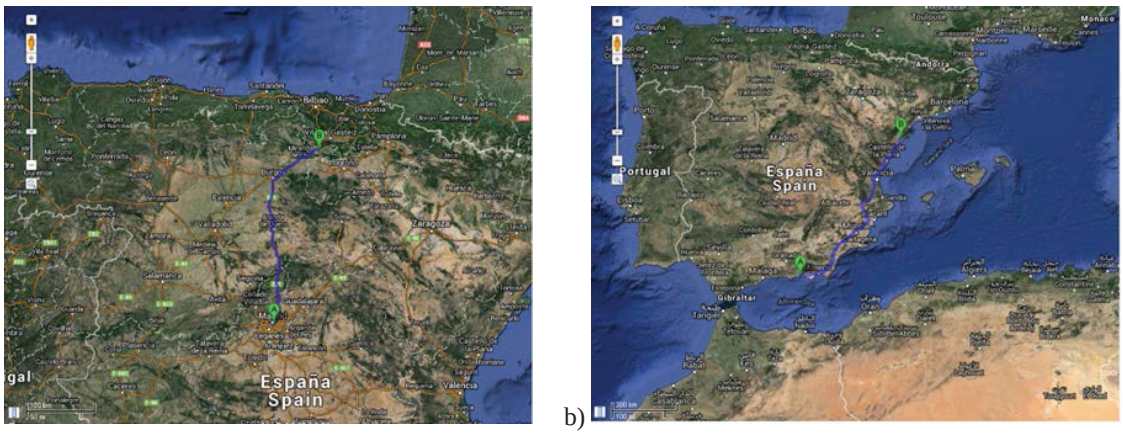


Fig. 1. a) Madrid-Irún route; b) Barcelona Almeria route.

The data used in this study come from two sources: police reports on Spanish accidents database (DB1) held by General Directorate of Traffic (DGT) and road geometry design measurements obtained from the Road Inventory database (DB2) from the General Directorate of Highways of the Ministry of Transportation (MFOM) (Table 1).

Table 1. Number of observations in 2 datasets.

Police accident reports (DB1)		Road geometry design (DB2)	
Year 2010	36,639	Madrid- Irún	78,001
Year 2011	33,004	Barcelona-Almería	20,889
Year 2012	32,589		

The dataset contains information on accidents that took place during 3 years (2010-2012) in two routes: Madrid-Irún and Barcelona-Almería and have a total of 3751 observations. The response variable, accident severity (ACSEV), describes whether the accident was severe or light. Severe accidents include fatalities as well.

The average annual daily traffic, AADT, in these axes, averaged over the three years, were: 28,700 vehicles in the Madrid- Irún and 29,300 in the Barcelona-Almeria route. Heavy traffic intensity is usually observed in high capacity road types (dual carriageways and motorways) than single carriageways segments of the routes.

Since the accident severity in different types of collisions is affected by different factors (Das et al., 2009, Harb et al., 2009)., the classification analyses were carried out for different accident types: 1) head-on collisions; 2) sideswipe-frontal collisions; 3) sideswipe; 4) rear-end collisions; and 5) collisions involving multiple vehicles as the result of a previous accident in the road segment. The percentage of each collision (cumulative percent) and the percentages of accident severity classes in each collision type are depicted in Fig.2.

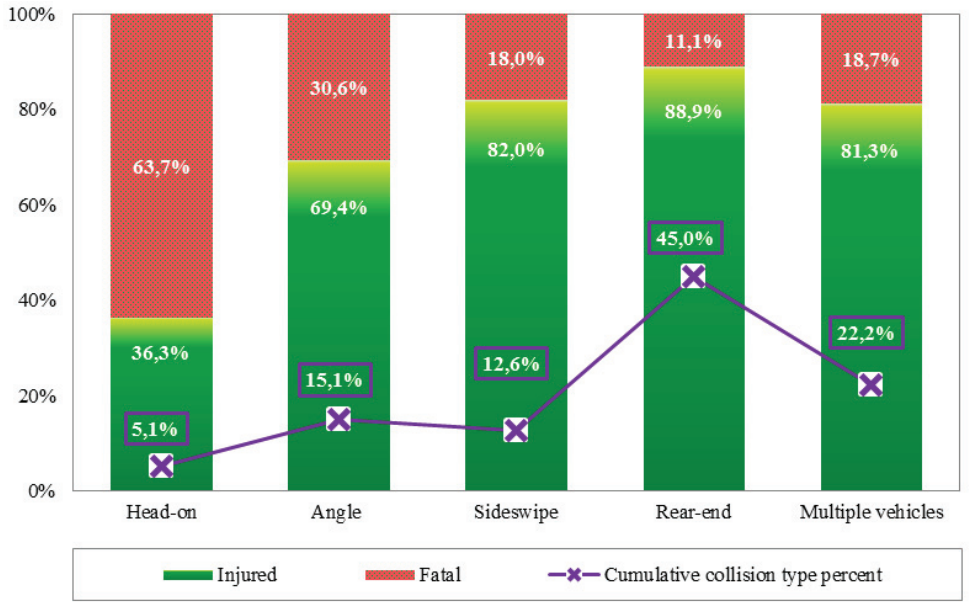


Fig. 2. Percentage of accident severity in each accident type.

The predictors used for the exploration of the accident severity are described in Table 2. There are a total of three road types in the above mentioned routes: double carriageways, single carriageways and toll roads. The traffic density in the road section at the time of the accident was classified as fluid, dense or congested. The alignment of the road section was divided into two categories: straight and curved. The surface of the road during the accident occurrence was categorized as follows: 1. dry and clean; 2. wet or humid; 3. frozen or snow-covered; and 4. greasy.

Visibility is represented by visibility obstruction, inverse and direct visibility. Visibility obstruction was used as a proxy variable to visibility level. As road geometry design standards, the number of lanes, main lane, median lane, shoulder lane and slow lane widths of the road sections, radius, superelevation and slope were considered. The data on shoulder and slow lane widths were initially available for both right and left road sections. In this study the total width of these road sections are considered.

Table 2. Description of the variables.

Variable	Name	Type	Description
Response variables:			
Accident severity	ACSEV	Binary	1. Accident resulted in at least one severe injury and/or fatality; 0. Accident resulted in at least one injury
Accident related factors (DB1):			
Traffic density	TD	Qualitative	1. Fluid; 2. Dense; 3. Congested
Alignment	AL	Qualitative	1. Straight; 2. Curved
Surface	SURF	Qualitative	1. Dry and clean; 2. Wet and humid;3. Frozen and snow covered; 4. Greasy
Visibility obstruction	VOB	Qualitative	1. Buildings; 2. Land shape; 3.Vegetation; 4. Atmospheric factors; 5. Sun glare; 6. Dust or smoke; 7. Other causes; 8. No obstruction
Geometry related factors (DB2):			
Road type	RTYP	Qualitative	1. Double carriageways; 2. Single carriageways; 3. Toll roads
Inverse visibility	INVS	Continuous	Inverse visibility range
Direct visibility	DRVS	Continuous	Direct visibility range
Number of lanes	NL	Count	Number of traffic lanes
Median width	MDW	Continuous	Width of the median; equal to 0 if there is no median
Main lane width	MLW	Continuous	Width of the main section lane
Shoulder width	SHW	Continuous	Sum of the left and right shoulder widths
Slow lane width	SLW	Continuous	Sum of the left and right slow lanes widths
Radius	RAD	Continuous	Radius of curvature
Superelevation	GRD	Continuous	Cross-platform tilt of the curve section.
Slope	SLP	Continuous	Slope of the road section

3. Random forests

The CART method which is based on the analysis of decision trees might become very unstable with the increasing number of observations and predictors. Therefore random forest methods (RF), have gained more attention recently for their ability to be become more stable and provide better predictions. The RF method was proposed by Breiman (2001) and is considered to be one the most efficient classification methods. RF method has garnered mostly favorable reviews when compared to logistic regression, quadratic discriminant analysis, support vector machines, classification and regression trees. It is based on Breiman's bagging principle of and Ho's random subspace that relies on constructing a collection of decision trees with random predictors.

Two main byproducts of RF are the out-of-bag error rate (OOB) and the variable importance. OOB is the misclassification rate which is computed after growing a tree with a bootstrapped sample (cluster).

The variable importance is measured through the classification accuracy and the Gini impurity rates. The importance rate of the variable in the forest is averaged over the number of grown trees:

$$VI(x_k) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(x_k)}{ntree} \quad (1)$$

where $VI(x_k)$ is the overall variable importance rate of the variable averaged over the trees across the RF, $ntree$; and $VI^t(x_k)$ is the importance rate of the variable at tree t.

4. Results and discussions

4.1. OOB error rate

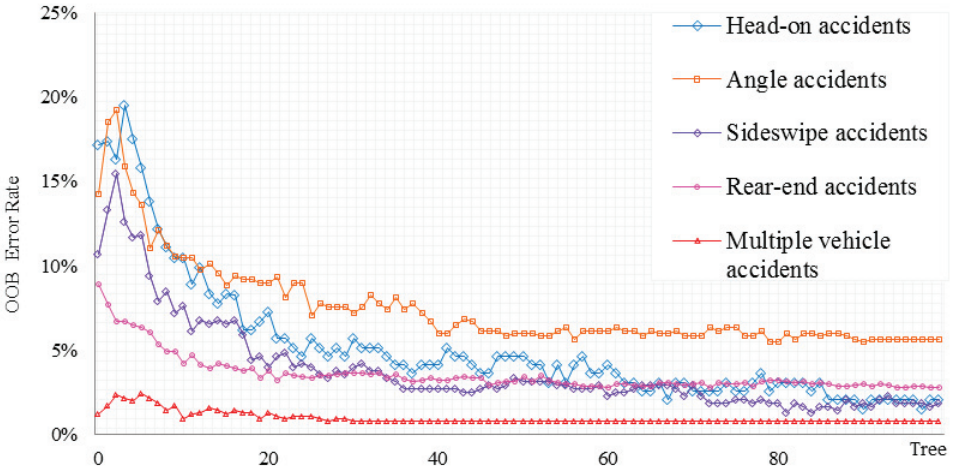


Fig. 3. Out-of-bag error rate.

In order to run RF, first the out-of-bag error rate of the trees (100) was computed. As it can be observed in Fig. 2, the OOB error rate becomes almost constant after 50 trees for each of the accident types. Therefore RF was run with 100 trees. The RF was first run by including all the variables. After the identification of the most important variables through Gini impurity, the RF was run for a second time in order to obtain information on the decision trees and the direction of the effects of important variables[1]. In the following sections the list of important variables affecting each accident type are discussed.

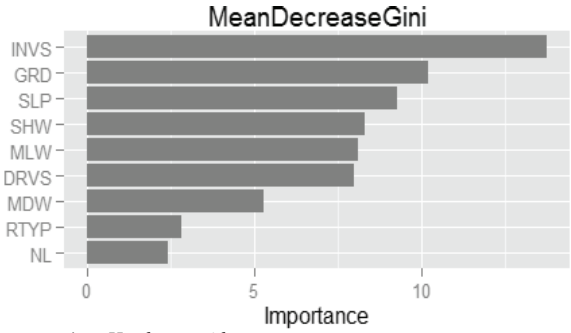
4.2. Variable importance

The importance rate of the variables by collision type according to Gini impurity measure are displayed in Fig. 3.

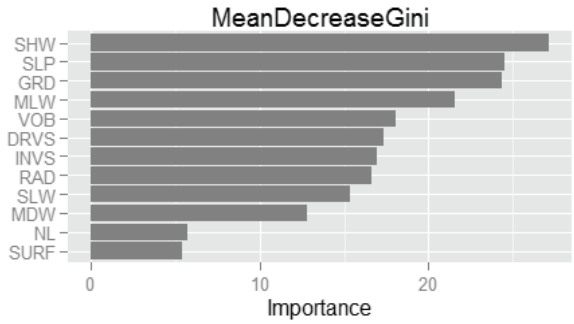
[1] Decision trees are not reported here but their results are briefly discussed in Conclusions.

4.2.1. Head-on accidents

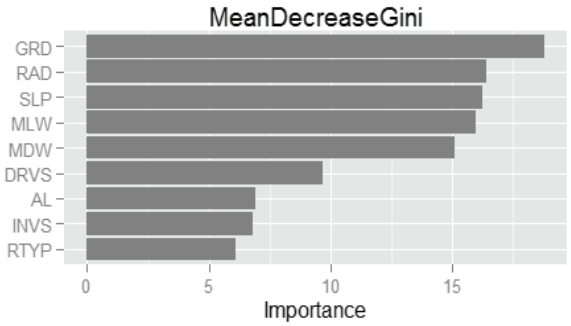
There were a total of 193 head-on accidents. 70 of these accidents 70 resulted in light injuries and 123 resulted in at least one fatality or severe injury. The RF was run once with all the variables to obtain the list of most important ones. The Gini impurity in Fig. 3 (a) shows the ranking of the variables according to their importance. The results show that the variables inverse visibility, direct visibility, superelevation, slope, main lane, shoulder lane and median lane widths are among the most important factors that contribute to head-on collisions and might be the defining factors in their outcome. The out-of-bag error rate of head-on collision random forests was 2.07%.



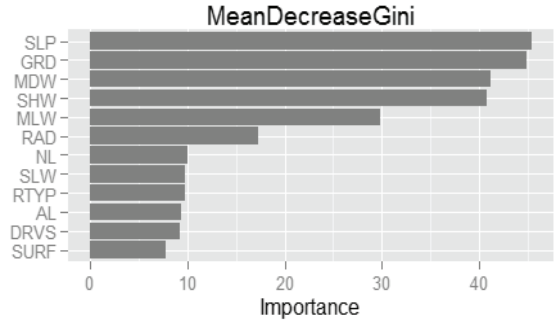
a) Head-on accidents



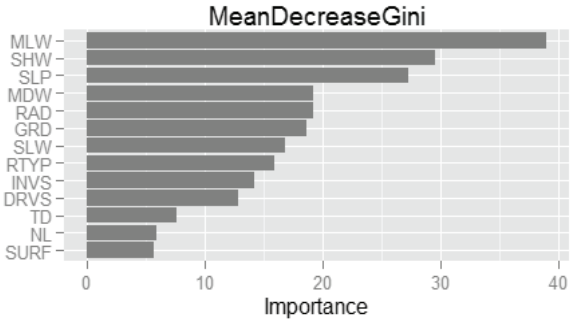
b) Angle accidents



c) Sideswipe accidents



d) Rear-end accidents



e) Multiple vehicle accidents

Fig. 4. Variable importance according to mean decrease Gini impurity.

4.2.2. Angle accidents

There were a total of 566 angle/turning movement accidents, out of which 393 resulted in light injuries while the remaining 173 resulted in either fatality or severe injury. The most important variables in this cluster, as shown in Fig. 3 (b) were shoulder lane and main lane widths, slope, superelevation, visibility obstruction and direct visibility. The OOB error rate of the random forest was 5.65%.

4.2.3. Sideswipe accidents

Out of 472 sideswipe accidents, 387 resulted in injury, while the remaining 85 resulted in at least one fatality or severe injury. The most important variables were radius, superelevation, main lane and median lane widths and direct visibility (see Fig. 3 (c)). The misclassification error rate of the sideswipe accidents RF was 1.91%.

4.2.4. Rear-end accidents

There were 1687 rear-end accidents recorded. 1500 of them resulted in light injuries while the remaining 187 resulted in fatality or severe injury. The most important variables affecting the severity of rear-end collisions WERE the slope, superelevation, radius, main lane, median lane, and slow lane and shoulder lane widths (see Fig. 3 (d). The misclassification error rate for the rear-end RF was 2.79%.

4.2.5. Multiple vehicle accidents

Out of 833 multiple accidents 677 resulted in injury while the remaining 156 resulted in at least one fatality or severe injury. The variables main lane, shoulder lane, slow lane and median lane widths, as well as slope, radius and the superelevation are among the most important predictors as far as the accident severity rate is concerned (see Fig. 3 (e)). The misclassification rate for this accident type was 0.84%.

5. Conclusions

Using random forests the severity rate of different accident types was analyzed. The accident data was collected during 3 years, 2010-2012 and includes all the accidents taking place during this time in Spain. The main objective of the study was to analyze the effect of geometry design standards on accident severity concerning the accidents taking place in 2 busy freight routes: Madrid-Irún and Almería-Barcelona.

The road geometry design is found to have a significant impact on the different accident types. Among the most important variables the main lane widths, superelevation and slope were found to affect the severity rate for all accident types.

The visibility range of the driver (direct and indirect visibility, visibility obstruction) was found to increase the severity of the head-on accidents, angle accidents and sideswipe accidents.

Other geometry design factors found to contribute to the severity of the accident were shoulder lane width (head-on, angle, rear-end and multiple vehicle accident types), median lane width (head-on, sideswipe and multiple vehicle accident types), slow lane width (rear-end and multiple vehicle accident types), and the radius (sideswipe, rear-end and multiple vehicle accident types).

The overall results of the study show that narrow main lane, shoulder lane, median lane and slow lane, might increase the accident severity. Higher superelevation and steeper slope also will increase the severity of the accident. These results indicate that the fatality or severe injury as a result of a road accident might in fact be prevented through the preliminary precautions by road design and planning operators. The results of this study could serve as a reference for future decision making by road operators.

Some aspects of this study might be limited. For example building shallow decision trees might cause a bias, although the results obtained through the decision trees mainly coincide with the importance ranking of the variables. This and other limitations of the work will be the focus of further research.

Acknowledgements

This work has been carried out in the framework of the MODALTRAM - TRA2011-28647-C02-01 Research Project "Development of an integrated methodology for the assessment of externalities (Safety and Environment) for the road and rail modal shift", of the Spanish National Research Plan 2011-2014, Ministry of Economy and Competitiveness (MINECO). The authors are thankful to General Directorate of Traffic (DGT) and General Directorate of Highways (DGC) of the Ministry of Transportation (MFOM) for the access to databases.

References

Abdel-Aty, M., Pande, A., Das, A., and Knibbe, W. (2008). Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. Transportation Research Record: Journal of the Transportation Research Board, (2083), 153-161.

Chang, L. Y., and Chen, W. C. (2005). Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research, 36(4), 365-375.

Das, A., Abdel-Aty, M., and Pande, A. (2009).Using conditional inference forests to identify the factors affecting crash severity on arterial corridors.Journal of safety research, 40(4), 317-327.

Das, A., and Abdel-Aty, M. (2010). A genetic programming approach to explore the crash severity on multi-lane roads. Accident Analysis & Prevention, 42(2), 548-557.

Harb, R., Yan, X., Radwan, E., and Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests. Accident Analysis & Prevention, 41(1), 98-107.

Ho, T. K. (1998). The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 20(8), 832-844.

Karlaftis, M. G., and Golias, I. 2002. Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis & Prevention,34(3), 357-365.

Kuhnert, P. M., Do, K. A., and McClure, R. (2000). Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis, 34(3), 371-386.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.